Glossary of terms used

Species and their phylogenetic relations

Species: groups of individuals that separated from other *species* (i.e. have *speciated*, hence *speciation*) so that they remain (at least temporarily) distinct at multiple divergent traits and genes across the genome, even when coexisting in the same areas (sympatry). Parts (and sometimes large fractions) of the genome, however, may remain undifferentiated because these regions exchange genes via *introgression* or *lateral transfer*. If this is so, it is assumed that the more divergent parts of the genome will contain the genetic differences that have contributed to speciation, and include loci, 'speciation genes,' that contribute to *reproductive isolation*. 'Speciation genes' are found not just in single co-inherited blocks, but are assumed to be scattered across the genome and to remain in very tight linkage disequilibrium among species in spite of gene exchange in genomic regions between these species-determining loci. Species are considered to continue to exist when there is *lateral transfer*, *hybridization* and *introgression*, provided that this is rare per generation compared to within-species exchange, so that the continued existence of separate species lineages is not in doubt. Over long periods of time large fractions, potentially over 50%, of the genome may be exchanged without threatening the existence of the separate species involved. The concept of species is highly controversial among some microbiologists (Woese 2002, Doolittle and Zhaxybayeva 2010, Lawrence and Retchless 2010), but others, while appreciating abundant *lateral transfer*, adopt the idea of species readily, if in a somewhat different way from those who work on eukaryotes (Hanage et al. 2005, Cohan 2010, Shapiro and Polz 2014).

Species phylogeny, or *species tree*: A tree that represents the bifurcation history and arrangement of *speciation* events of extant species and their ancestors. Reticulation events that have occurred since *speciation* are presumed not to have affected the majority of genes involved in maintaining species, although the fraction of the genome showing the true species tree may be very small. Others in the past have disagreed with this, arguing that there may be no grounds for a *species tree* separate from the overall average or 'democratic' signal of the gene trees: "In considering these issues [incomplete lineage sorting, lateral transfer, introgression], one is provoked to reconsider precisely what is [*the true*] phylogeny. Perhaps it is misleading to view some gene trees as agreeing and other gene trees as disagreeing with the *species tree*; rather, all of the gene trees are part of the *species tree*, which can be visualized like a fuzzy statistical distribution, a cloud of gene histories" (Maddison 1997). Unfortunately for this view, recent theoretical studies of *incomplete lineage sorting* show that in some rapidly branching trees (the 'anomaly zone'), the maximum likelihood or 'democratic' signal of gene trees may diverge from the true *species tree* (Degnan and Rosenberg 2009). If we fail to accept the origin and continuation of recognizably separate entities (species) as contributing to a tree - in spite of *incomplete lineage sorting*, *lateral transfer*, and *introgression* - then there is no basis at all for talking about a *species tree*. If these processes lead to substantial *admixture* among even species-determining loci (and this may well happen, sometimes, for example in *hybrid speciation*), then there will be no true recoverable tree of species at all: the species phylogeny may not be a tree, and indeed, all of life could be regarded asreticulate and below the Darwinian Threshold of Woese (Woese 2002).

Comment [MH1]: I'm not sure I re-phrased this in the best way, but "democratic" approaches to "determining" species trees is a different issue.

The 'true' phylogeny: *The true phylogeny* represents the true history of all of the genes of the species as well as complex and quantitative measures of gene exchange for different regions of the genome. It is clearly best represented as a net rather than a tree; that is, the *true phylogeny* will actually be reticulate. As pointed out in the main text, this is more or less enforced by the invention of meiosis by eukaryotes, by the existence of some sexual exchange among species, and by high levels of lateral transfer in some parts of the Tree of Life. A reticulate *true phylogeny* does not preclude the existence of a *species phylogeny* that is a tree, provided that *lateral transfer*, *hybridization* and *introgression* are not so pervasive as to blend two branches into a single cluster at some point in time after speciation.

Monophyletic, *paraphyletic, polyphyletic*: The usual meanings, but we use *monophyletic* to mean taxa in a *species tree* that include all descendants of an ancestor that had a single origin and have persisted as recognizably separate entities since their origins, potentially in spite of much *reticulation* with other groups since that origin. Under this scheme, for example, Eukaryota, Plants, and Animals are considered putatively *monophyletic* in spite of clear evidence for horizontal transfer of multiple genes with other eukaryotes and with prokaryotes. A more elaborate classification of phylogenetic groupings with reticulate *true phylogenies* has been presented, adding the terms *epiphyletic*, *periphyletic*, and *anaphyletic* to existing terms (Wheeler 2014). We do not use these terms here.

Genealogical reticulation or genealogical incongruence: Different parts of the genome may differ from each other and the *species tree*, for a variety of reasons: *incomplete lineage sorting* (*ILS*), lateral gene transfer (*LGT*), or *introgression*.

Incomplete lineage sorting, *lineage sorting* (*ILS*): Failure of a gene genealogy to coalesce within a tree branch representing a species. Coalescence in an ancestral species can lead to *genealogical reticulation*, even though neither the *species tree* nor the *true phylogeny* is reticulate, i.e. there is no *lateral transfer* or *introgression* to disturb the tree.

Reticulation, phylogenetic incongruence: involves *LGT* or *introgression*, so that a true representation of the *species phylogeny* is reticulate. *ILS* is not counted as 'true' *phylogenetic reticulation* or *incongruence*.

Hybridization: Equivalent to *m* or gene flow, measured as the fraction of first generation (F_1) hybrids produced by each species per generation.

Introgression: Equivalent to 'successful' backcrossing by the F1 hybrids. Evidence of *introgression* may be transient, due to countervailing selection. Overlaps somewhat with *admixture*.

Admixture fraction: Permanent effect of *gene flow*. This term is in use, especially among human geneticists (Green et al. 2010), to refer to the fraction of the genome that originated from another species (or population). *Introgressed fraction* is a synonym used by others, so *admixture* overlaps somewhat with *introgression*.

Components of reproductive isolation

Comment [MH2]: ow I am confused. I thought this is the definition of “species phylogeny.”

Comment [MJ3]: You were right. It took me a long time to realize it, but I had blurred my emergent idea of "true phylogeny" and "species phylogeny" in both my mind and this section of the text. See if you like this version. And anyway, what do you think? I feel we are actually in new territory here.

Comment [MH4]: No, I am becoming more convinced that we have these backwards, and possibly even that admixture should be avoided altogether. It is clear in the literature that admixture can be applied to populations of the same species, but that introgression can never be used this way. And “admixture mapping” is explicitly about polymorphism still segregating in populations, regardless of the number of samples used in Patterson’s tests.

Comment [MJ5]: We agree to disagree on this.

Gene flow (between species): Actual flow of genes between *species*. Note, this is intended to be *gene flow* before the action of selection to remove introgressed variants. Equivalent to *hybridization* (above). In contrast, in some of the speciation literature, *'gene flow*' can refer to *effective gene flow* that has a similar meaning to our *introgression* or *admixture* as defined above.

Assortative mating (or assortative fertilization): Behavioural or gametic interactions that lead to a reduction in mating or fertilization rates, and hence to a reduction in *hybridization* or *gene flow* between populations. Often referred to in the speciation literature as *'prezygotic isolating mechanisms'* or *'prezygotic isolation.'*

Divergent selection: A process that leads to low fitness of introgressed alleles. This might include epistatic selection against certain gene combinations (Dobzhansky-Muller incompatibilities), direct selection against alleles that are unfavourable in the environment of the new species ('immigrant inviability' (Nosil et al. 2005)), and diversifying or disruptive phenotypic selection (Schluter 2001, Coyne and Orr 2004). May lead to *'postzygotic isolating mechanisms*,*' 'postzygotic isolation,' or Dobzhansky-Muller incompatibilities*, which can be either intrinsic (genomic effects on the individual) or extrinsic (affect ecological interactions of the individual with its environment, including other species).

Reproductive isolation: A quality that species are generally supposed to have. This is a commonly used term in the speciation literature that can be prone to misinterpretation. When estimated quantitatively, it can measure either the lack of *hybridization* or the lack of permanent *introgression* among species, although there's no consensus of exactly how to do this, and different methods give different answers (Sobel and Chen 2014). In common parlance, reproductive isolation consists of any combination of '*prezygotic'* and '*postzygotic isolation*,' depending on context. In other words, reproductive isolation as a whole represents an 'effective' measure of gene flow that represents the balance between *divergent selection* and *gene flow*.

Dobzhansky-Muller incompatibilities: Divergent genetic loci between species that alone have no deleterious effects on hybrids, but in concert cause epistatic incompatibilities. They are liable to evolve between populations or species with weak or no gene flow contact, because natural selection prohibits their polymorphism, while there is no barrier to the evolution of the divergence between species. They are regarded as a form of '*postzygotic isolation.*'

Forms of gene exchange among species

Horizontal (or lateral) gene transfer (LGT): In the current paper, we use the term *lateral transfer* to mean a form of *non-meiotic exchange*, potentially between very distant taxa. In the literature, *lateral transfer* can refer to the effects of any process, including *sex* that leads to transfer of genes among species or populations. For instance, conjugation and transfer of plasmids between strains of *E.coli* is considered a form of horizontal transfer by some microbiologists (Redfield 2001). We here include these, as well as certain types of prokaryotic *homologous exchange*, such as transformation among related strains or species (Zawadzki et al. 1995), as components of *sex*, rather than LGT. Although the two clearly overlap, this is an attempt to bring prokaryote sex vs. horizontal transfer closer to what we mean by sex vs. horizontal transfer in eukaryotes.

Sex: Here we use sex to mean 'regular,' usually homologous genetic exchange among closely related organisms, often within the same species, to distinguish it from LGT. Others have used alternative definitions of sex for prokaryotes that include LGT (Redfield 2001).

a. *Prokaryotic sex* vs. LGT is not well defined, but usually involves homologous exchange of genes. Mechanisms of prokaryote sex include conjugation, transduction, and transformation. These processes tend to be most successful among closely related strains (especially within *species*). Cellular fusion leading to recombination of large parts of the genome has recently been described between different populations of culturable Archaea (Naor et al. 2012), but this does not lead to reciprocal exchange, as in meiosis. Prokaryote sex can be characterized for the most part as the acquisition of a few new genes by a genome from a related source. Prokaryote sex appears never to be reciprocal, where progeny cells may have roughly equal fractions of each parental genome.
b. In contrast, *eukaryotic sex* always involves cell fusion and meiosis that generally leads to *homologous* or *reciprocal* exchange across the whole genome. In meiosis each daughter haploid product acquires a recombined ~50% of the genome from each parent haploid. Diploid eukaryotes consist of organisms with genomes originally formed via haploid cell fusion; they contain homologous copies of both parental genomes.

Non-homologous exchange: a transfer process whereby genes are added to a genome, rather than being replaced by a homologous copy. Acquisition of new genes by bacterial genomes via LGT, for example, is an example of non-homologous exchange.

Homologous exchange: a transfer process whereby existing genes in the genome are replaced by a related copy from another individual, including between species. Homologous exchange between bacterial cells is considered, for the purposes of this paper, to be close to what we mean by bacterial *sex*.

Classification

Names of taxa listed below are used informally. We adopt terms that we hope are generally in use and are understood to describe particular groups of species for communication purposes. Our use of these names, or of the names of any other taxon, does not imply belief in monophyly or lack of reticulation of the group indicated. Instead, the groups these may be monophyletic, paraphyletic, or even polyphyletic (at least if due to gene transfer).

Prokaryotes: Lack nucleus, mitochondria, nuclear membrane, and meiosis. A paraphyletic reticulate group consisting of *Bacteria* and *Archaea*. The group is believed to be paraphyletic because the root of the tree of life (on the basis of a few genes such as 16S RNA) is said to separate the *Bacteria* from the *Archaea+Eukaryota*

Archaea, Archaebacteria: Prokaryotes distinguished originally from Bacteria via 16S RNA sequences. A heterogeneous group of prokaryotes more closely related to eukaryotes at informational genes (i.e. those involved in transcription and translation) than either are to Bacteria (Woese and Fox 1977, Woese 2002). The Archaea are today believed very likely paraphyletic relative to eukaryotes (Williams and Embley 2014).

Bacteria, Eubacteria: Prokaryotes were originally distinguished from Archaea via 16S RNA sequences. A heterogeneous group of prokaryotes more distantly related to eukaryotes at informational genes than Archaea.

Eukaryotes, Eukaryota, Eukarya: A putatively monophyletic group (although much horizontal transfer is evident with prokaryotes), closer to the Archaea than to the Bacteria at informational genes in the three-domain system. In eukaryotes, the genome is enclosed in a nuclear envelope, and the cytoplasm typically has mitochondria. Meiosis is a characteristic of eukaryote sexual reproduction, and is absent in prokaryotes.

Protists: Diverse paraphyletic group consisting of single-celled eukaryotes, excluding fungi (which include single-celled microsporidia, yeasts, etc.), plants, and animals.

Plants: Putatively monophyletic autotrophs that include multicellular organisms containing chloroplasts (derived from cyanobacterial endosymbiosis), and with cellulose cell walls. Thought to be derived from a green alga-like (Chloroplastida) single-celled protist ancestor (Adl et al. 2012).

Fungi: Putatively monophyletic group of unicellular and multicellular heterotrophs lacking chloroplasts, but with cell walls containing chitin. Thought to be derived from a nucleariid (nuclearian)-like filose amoeba ancestor (Adl et al. 2012).

Animals, Metazoa: Putatively monophyletic multicellular heterotrophs, generally motile or with some motile stages. Animals lack chloroplasts (except via endosymbiosis with photosynthetic protists) or cell walls. Thought to be derived from a chaonoflagellate (chaonozoan)-like protist ancestor (Adl et al. 2012).

Adl SM, et al. 2012. The revised classification of eukaryotes. Journal of Eukaryotic Microbiology 59:429-514.
Cohan FM. 2010. Are species cohesive? A view from bacteriology. Pages 000- in Walk S, Feng P, eds. Washington, DC: American Society for Microbiology Press.
Coyne JA, Orr HA. 2004. Speciation. Sinauer Associates.
Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology and Evolution 24:332-340.
Doolittle WF, Zhaxybayeva O. 2010. Metagenomics and the units of biological organization. BioScience 60:102-112.
Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.
Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3:6.
Lawrence JG, Retchless AC. 2010. The myth of bacterial species and speciation. Biology and Philosophy 25:569-588.
Maddison WP. 1997. Gene trees in species trees. Systematic Biology 46:523-536.
Nosil P, Vines TH, Funk DJ. 2005. Reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59:705-719.
Redfield RJ. 2001. Do bacteria have sex? Nature Reviews Genetics 2:634-639.
Schluter D. 2001. Ecology and the origin of species. Trends in Ecology and Evolution 16:372-380.
Shapiro BJ, Polz MF. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology 22:235-247.
Sobel JM, Chen GF. 2014. Unification of methods for estimating the strength of reproductive isolation. Evolution 68:1511-1522.
Wheeler WC. 2014. Phyletic groups on networks. Cladistics 30:447-451.
Williams TA, Embley TM. 2014. Archaeal “dark matter” and the origin of eukaryotes. Genome Biology and Evolution 6:474-481.
Woese CR. 2002. On the evolution of cells. Proceedings of the National Academy of Sciences of the United States of America 99:8742-8747.
Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74:5088-5090.
Zawadzki P, Roberts MS, Cohan FM. 1995. The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* is robust. Genetics 140:917-932.